Whole Sentence Neural Language Models

*Yinghui Huang, Abhinav Sethy *, Kartik Audhkhasi, Bhuvana Ramabhadran **

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

* Work was done in IBM

Outline

- Introduction
  - Conditional models
  - Whole sentence language models
- Training
  - Noise Contrastive Estimation (NCE)
  - Sampling for NCE
- Experimental results
  - Sequence Identification tasks
  - Large Vocabulary Automatic Speech recognition
- Conclusion

Outline

- Introduction
  - Conditional models
  - Whole sentence language models
- Training
  - Noise Contrastive Estimation (NCE)
  - Sampling for NCE
- Experimental results
  - Sequence Identification tasks
  - Large Vocabulary Automatic Speech recognition
- Conclusion

Background

- Most of the statistical LMs are conditional models

$$p(s) = p(w_1, \dots, w_T) = \prod_{t=1}^{T} p(w_t | w_{t-n+1}, \dots, w_{t-1})$$

Background

- Most of the statistical LMs are conditional models

  $p(s) = p(w_1, \dots, w_T) = \prod_{t=1}^{T} p(w_t | w_{t-n+1}, \dots, w_{t-1})$

- The locally-conditional design limits the ability of the model in exploiting whole sentence structure.
  - It makes implicit independence assumptions that may not be always true.
  - Global sentence information may be difficult to capture.
- Whole sentence maximum entropy LMs directly models $p(s)$, probability of a sentence or a utterance. (Rosenfeld, 1997; Chen, 1999)

Background

- Most of the statistical LMs are conditional models

$$p(s) = p(w_1, \dots, w_T) = \prod_{t=1}^{T} p(w_t | w_{t-n+1}, \dots, w_{t-1})$$

- The locally-conditional design limits the ability of the model in exploiting whole sentence structure.
  - It makes implicit independence assumptions that may not be always true.
  - Global sentence information may be difficult to capture.
- Whole sentence maximum entropy LMs directly models $p(s)$, probability of a sentence or a utterance. (Rosenfeld, 1997; Chen, 1999)
- Recurrent neural network (RNN) LMs are proposed to capture longer histories (Hochreiter, 1997; Mikolov, 2010)

Whole sentence Neural Language Models

- We combine whole-sentence LMs with LSTM.

Whole sentence Neural Language Models

- We combine whole-sentence LMs with LSTM.

- Initial attempt

$$p(s) = \mathrm{softmax}\big(f(s)\big) = \frac{1}{Z} \cdot \exp\big(f(s)\big)$$

Whole sentence Neural Language Models

- We combine whole-sentence LMs with LSTM.

- Initial attempt

$$p(s) = \text{softmax}\big(f(s)\big) = \frac{1}{Z} \cdot \exp\big(f(s)\big) = \frac{1}{\sum_{s\prime} \exp(f(s\prime))} \cdot \exp\big(f(s)\big)$$

- Calculating Z is infeasible, since it involves summing all possible sentence s.

- To train an un-normalized model instead !

Whole sentence Neural Language Models

- We combine whole-sentence LMs with LSTM.
- We use Noise Contrastive Estimation (Gutmann, 2012) for training to avoid normalization.

- Our model
  - does not require any softmax computation to compute conditional probabilities of individual words.
  - generates a single output score for the whole sentence which we treat as an un-normalized probability.

Outline

- Introduction
  - Conditional models
  - Whole sentence language models
- Training
  - Noise Contrastive Estimation (NCE)
  - Sampling for NCE
- Experimental results
  - Sequence Identification tasks
  - Large Vocabulary Automatic Speech recognition
- Conclusion

Noise Contrastive Estimation

- NCE was first introduced as a sampling-based approach for unnormalized training of statistical models. (Gutmann, 2012)
- It has been widely used for improving the scalability of conditional neural net based LMs. (A. Mnih, 2012; Chen, 2017)
- With sufficient samples, the model learns the data distribution, also implicitly constrains the normalization term to be 1.

Sampling for NCE

- We use back-off n-gram LMs built on the training data as noise samplers
- Two types of noise samples:
  1. Generate word sequences using noise sampler model (RAND).

     Example: *March the twenty fifth of March nineteen twenty thirteen*
  2. Sample from an edit transducer (Mohri 2002). We first randomly select one sentence from the training data, then randomly select N positions to introduce an insertion (INS), substitution (SUB) or deletion (DEL) error.

REF	July the twentieth nineteen seventy nine
SUB	July twenty twentieth nineteen seventy nine
INS	July the twentieth nineteen ninety seventy nine
DEL	July the twentieth * seventy nine

Outline

- Introduction
  - Conditional models
  - Whole sentence language models
- Training
  - Noise Contrastive Estimation (NCE)
  - Sampling for NCE
- Experimental results
  - Sequence Identification tasks
  - Large Vocabulary Automatic Speech recognition
- Conclusion

Sequence Identification tasks

- Proof of concept: validate the idea that the model can detect patterns relying on entire sentence structures.
- Data
  - Palindrome.
    - 1M-word corpus with 10-word vocabulary.
    - Example: *the cat ran fast ran cat the*
  - Lexicographically-ordered words.
    - 1M-word corpus with 15-word vocabulary.
    - Example: *bottle cup haha hello kitten that what*
  - Expressing dates.
    - 7M-word corpus with a 70-word vocabulary.
    - Example: *January first nineteen oh one*

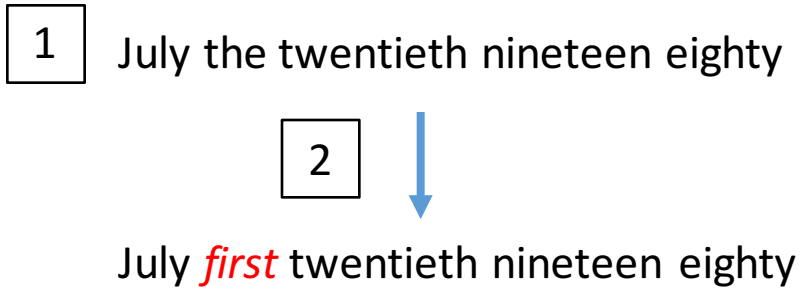
Sequence Identification tasks

- Task
  1. 10% of the generated data was used as the test set.
  2. Imposter sentences are generated by substituting one word
  3. Scores are assigned by the model for each sentence. A binary linear classifier will be trained to classify these scores into two classes.
  4. Performance is evaluated by its classification accuracy.

1	July the twentieth nineteen eighty

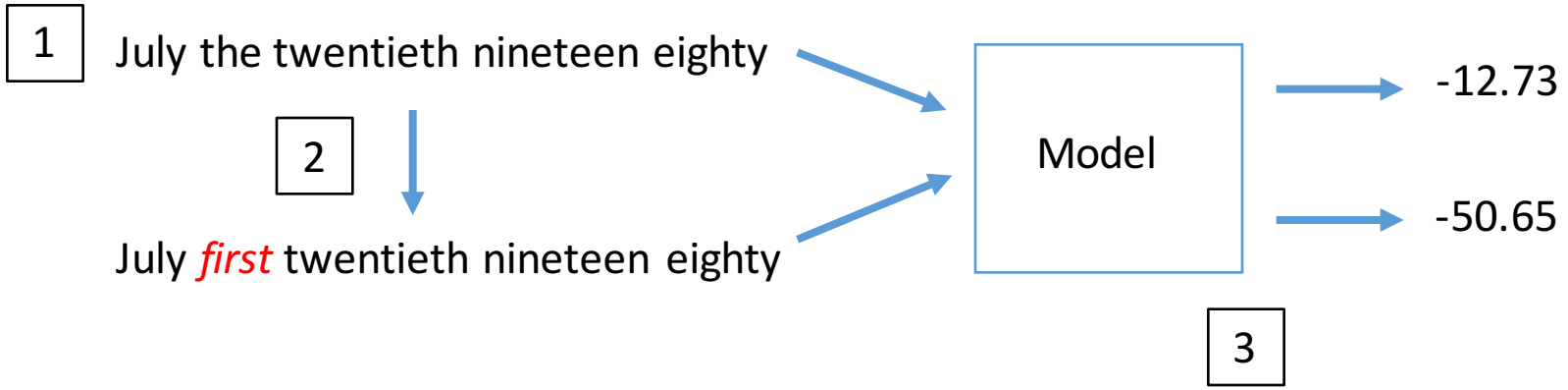
Sequence Identification tasks

- Task
  1. 10% of the generated data was used as the test set.
  2. Imposter sentences are generated by substituting one word
  3. Scores are assigned by the model for each sentence. A binary linear classifier will be trained to classify these scores into two classes.
  4. Performance is evaluated by its classification accuracy.



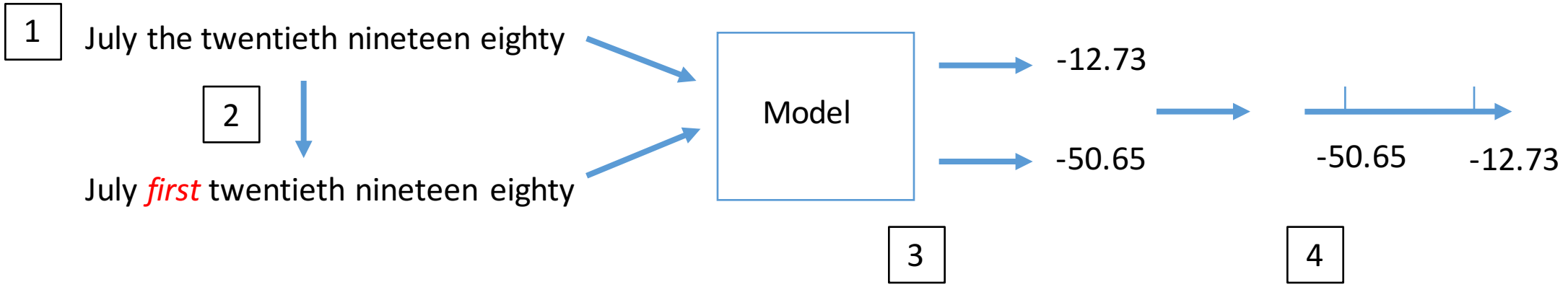
Sequence Identification tasks

- Task
  1. 10% of the generated data was used as the test set.
  2. Imposter sentences are generated by substituting one word
  3. Scores are assigned by the model for each sentence. A binary linear classifier will be trained to classify these scores into two classes.
  4. Performance is evaluated by its classification accuracy.



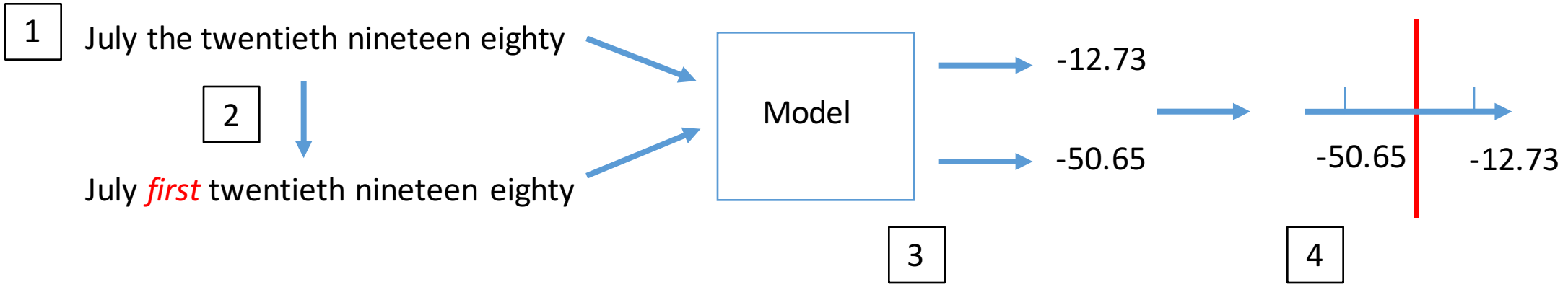
Sequence Identification tasks

- Task
  1. 10% of the generated data was used as the test set.
  2. Imposter sentences are generated by substituting one word
  3. Scores are assigned by the model for each sentence. A binary linear classifier will be trained to classify these scores into two classes.
  4. Performance is evaluated by its classification accuracy.



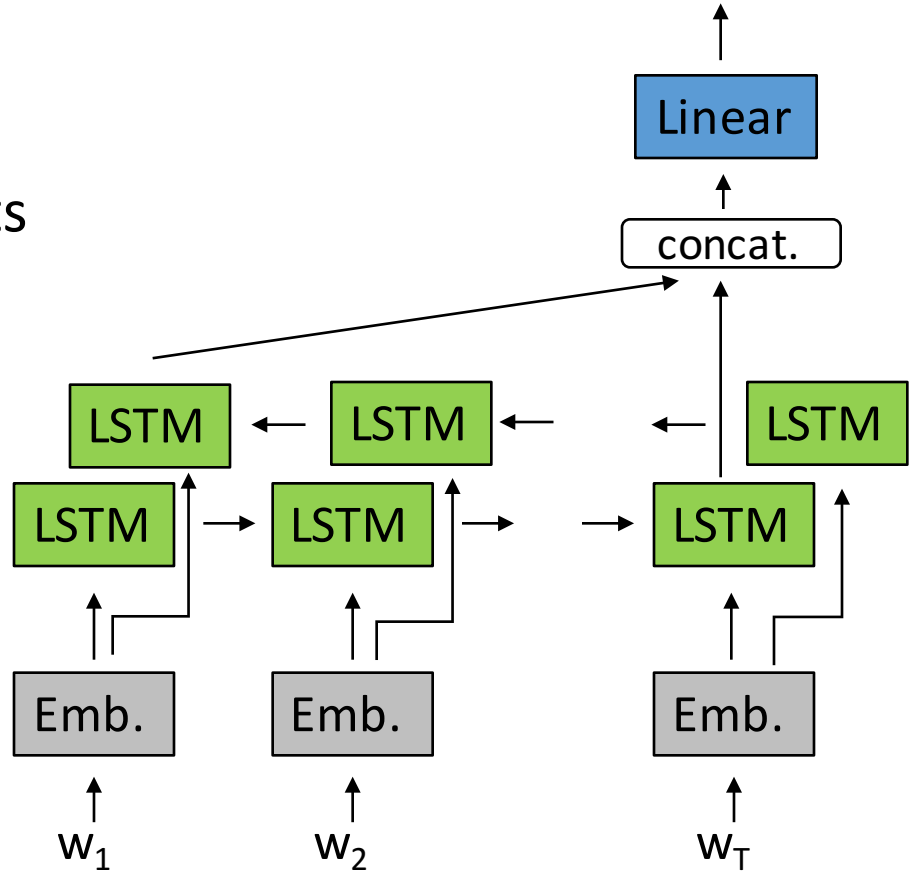
Sequence Identification tasks

- Task
  1. 10% of the generated data was used as the test set.
  2. Imposter sentences are generated by substituting one word
  3. Scores are assigned by the model for each sentence. A binary linear classifier will be trained to classify these scores into two classes.
  4. Performance is evaluated by its classification accuracy.



Experimental Results – Sequence Identification tasks

- Model configuration
  - One-layer BiLSTM
  - Embedding size of 200 with 700 hidden units



Experimental Results – Sequence Identification tasks

- For all three tasks, accuracy on average is above 99%.

Experimental Results – Sequence Identification tasks

- A closer look on DATE test set

Table 1: Example sentences from the DATE test set

REF	July the twentieth nineteen seventy nine
SUB	July twenty twentieth nineteen seventy nine
INS	July the twentieth nineteen ninety seventy nine
DEL	July the twentieth * seventy nine
RAND	July the twenty seventh of September two thousand eighteen

Table 2: Classification error rate (%)

	4-gram	sentence model
REF	0.03	0.00
SUB	0.73	0.04
INS	0.01	0.00
DEL	2.22	0.00
RAND	22.70	0.40

Experimental Results – Sequence Identification tasks

- We hypothesize that it is because the sentence model does not make conditional independence assumptions inherent in the locally-conditional models.

Table 1: Example sentences from the DATE test set

REF	July the twentieth nineteen seventy nine
SUB	July twenty twentieth nineteen seventy nine
INS	July the twentieth nineteen ninety seventy nine
DEL	July the twentieth * seventy nine
RAND	July the twenty seventh of September two thousand eighteen

Table 2: Classification error rate (%)

	4-gram	sentence model
REF	0.03	0.00
SUB	0.73	0.04
INS	0.01	0.00
DEL	2.22	0.00
RAND	22.70	0.40

Experimental Results – Speech Recognition

- Test set
  - Hub5 Switchboard-2000 benchmark task (SWB)
  - In-house Conversational Interaction task (CI)

    Test set is of duration 1.5 hours, consisting of accented data covering spoken interaction in concierge and other similar application domains.

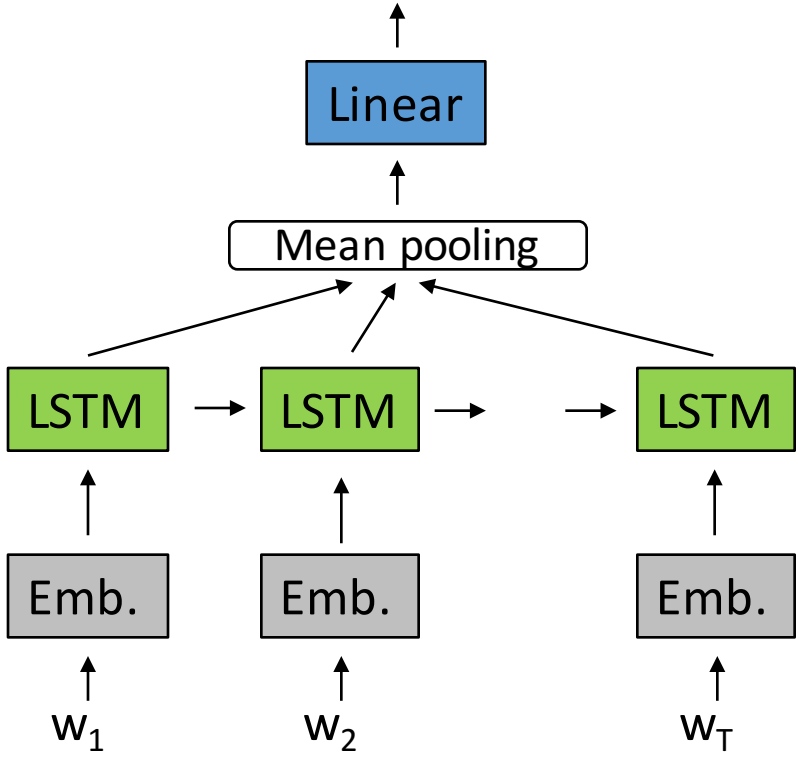
    Examples:
    - *can i request a room on a lower floor*
    - *is there a charge to use the fitness room*

- Evaluation on N-best (N=100) list rescoring.

Experimental Results – Speech Recognition

- Model configuration
  - One layer uni-directional LSTM

Test set	Projection layer	Hidden layer
SWB	512	512
CI	256	256



- NCE Noise samples drawn from 1-edit to 3-edit

Experimental Results – Speech Recognition

- Rescoring result (word error rate %)

	SWB	Conversational Interaction
N-gram	6.9	8.5
+ word LSTM	6.5	8.5
+ sentence Model	6.3	8.3

Experimental Results – Speech Recognition

- Rescoring result (word error rate %)

	SWB	Conversational Interaction
N-gram	6.9	8.5
+ word LSTM	6.5	8.5
+ sentence Model	6.3	8.3

- Example

Reference	actually we were looking at the saturn S L two
N-gram LM	actually we were looking at the *saturday I sell to*
+ Word LSTM	actually we were looking at the *saturday* S L *too*
+ Sentence LM	actually we were looking at the saturn S L *too*

Experimental Results – Speech Recognition

- Rescoring result (word error rate %)

	SWB	Conversational Interaction
N-gram	6.9	8.5
+ word LSTM	6.5	8.5
+ sentence Model	6.3	8.3

- Example

Reference	Could you send some soda to room three four five
N-gram LM + word LSTM	Could you send some *sort of* to room three four five
+ Sentence LM	Could you send some soda to room three four five

Outline

- Introduction
  - Conditional models
  - Whole sentence language models
- Training
  - Noise Contrastive Estimation (NCE)
  - Sampling for NCE
- Experimental results
  - Sequence Identification tasks
  - Large Vocabulary Automatic Speech recognition
- Conclusion

Conclusion

- We propose whole sentence neural language models, which estimates the probability for the entire word sequence directly with LSTM.
- To avoid normalizing over the whole sentence space, we apply NCE for training our recurrent nets.
- The preliminary results on a range of tasks show that the model captures information out from locally-conditional constraints.
- The proposed approach can be extended to other neural network architectures.

Reference

- Tomas Mikolov, Martin Karafi´at, Luk´as Burget, Jan Cernock´y, and Sanjeev Khudanpur, "Recurrent neural network based language model.," in Interspeech, 2010, vol. 2, p. 3.
- S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
- Le Hai Son, Alexandre Allauzen, and Franc¸ois Yvon, "Measuring the influence of long range dependencies with neural network language models," in Proceedings of the NAACLHLT 2012 Workshop: Will We Ever Really Replace the Ngram Model? On the Future of Language Modeling for HLT, Stroudsburg, PA, USA, 2012, WLM '12, pp. 1–10, Association for Computational Linguistics.
- Ronald Rosenfeld, "A whole sentence maximum entropy language model," in Proceedings of the IEEEWorkshop on Speech Recognition and Understanding, 1997.
- Stanley F. Chen and Ronald Rosenfeld, "Efficient sampling and feature selection in whole sentence maximum entropy language models," in ICASSP, Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, 1999.
- Yoon Kim, "Convolutional neural networks for sentence classification," in EMNLP. 2014, pp. 1746–1751, ACL.
- Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Skip-thought vectors," CoRR, vol. abs/1506.06726, 2015.
- Mingbo Ma, Liang Huang, Bing Xiang, and Bowen Zhou, "Dependency-based convolutional neural networks for sentence embedding," arXiv preprint arXiv:1507.01839, 2015.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu, "Towards universal paraphrastic sentence embeddings," in Proceedings of ICLR, 2016.
- Bo Pang and Lillian Lee, "A sentimental education: Sentiment analysis using subjectivity," in Proceedings of ACL, 2004, pp. 271–278.
- Brian Roark, Murat Saraclar, Michael Collins, and Mark Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 47.
- Erinc¸ Dikici, Murat Semerci, Murat Sarac¸lar, and Ethem Alpaydin, "Classification and ranking approaches to discriminative language modeling for asr," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 291–300, 2013.

- Libin Shen, Anoop Sarkar, and Franz Josef Och, "Discriminative reranking for machine translation.," in HLT-NAACL, 2004, pp. 177–184.
- Michael U Gutmann and Aapo Hyv¨arinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," The Journal of Machine Learning Research, vol. 13, no. 1, pp. 307–361, 2012.
- Andriy Mnih and Yee Whye Teh, "A fast and simple algorithm for training neural probabilistic language models," in Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 1751–1758.
- Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang, "Decoding with large-scale neural language models improves translation.," in EMNLP. Citeseer, 2013, pp. 1387– 1392.
- Abhinav Sethy, Stanley Chen, Ebru Arisoy, and Bhuvana Ramabhadran, "Unnormalized exponential and neural network language models," in ICASSP, 2015.
- Xie Chen, Xunying Liu, Mark J. F. Gales, and Philip C.Woodland, "Recurrent neural network language model training with noise contrastive estimation for speech recognition," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5411–5415, 2015.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 3111–3119. Curran Associates, Inc., 2013.
- Mehryar Mohri, "Edit-distance of weighted automata," in CIAA. Springer, 2002, vol. 2, pp. 1–23.H.
- Bourlard and N. Morgan, "Generalization and parameter estimation in feedforward nets: Some experiments," in Advances in Neural Information Processing Systems, 1990, vol. II, pp. 630–637.
- L. Brandchain, "The mixer 6 corpus: Resource for crosschannel and text independent speaker recognition," LREC, 2010.
- Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," Language Resources and Evaluation, vol. 41, no. 1, pp. 181–190, 2007.
- S Itahashi, "Recent speech database projects in japan," Proc. ICSLP, 1990.
- Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada, "Acoustical sound database in real environments for sound scene understanding and handsfree speech recognition.," in LREC, 2000.
- George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al., "English conversational telephone speech recognition by humans and machines," arXiv preprint arXiv:1703.02136, 2017.

Whole Sentence Neural Language Models

*Yinghui Huang, Abhinav Sethy *, Kartik Audhkhasi, Bhuvana Ramabhadran **

IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

* Work was done in IBM